On permutation symmetries in Bayesian neural network posteriors: a variational perspective

Simone Rossi[*], Ankit Singh[§], Thomas Hannagan[*]

[*]Stellantis (France), [§]Stellantis (India)

Motivating observation

Observation: Neural networks have many symmetries that are functionally equivalent. Recent evidence that SGD solutions are linearly connected if we account for permutations symmetries.

Ainsworth, Samuel, Hayase, Jonathan, and Srinivasa, Siddhartha. 2023.
Entezari, Rahim et al. 2022.

Motivating observation

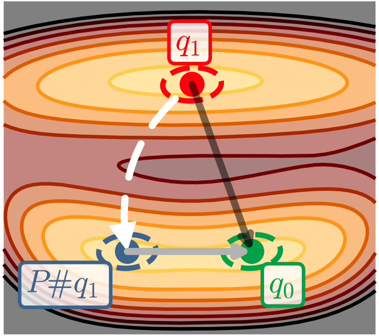
STELLANTIS

Observation: Neural networks have many symmetries that are functionally equivalent. Recent evidence that SGD solutions are linearly connected if we account for permutations symmetries.

Question: Do BNNs (and variational inference) share the same linearly connected behavior after accounting for functionally equivalent permutations?

Conjecture: Yes

Log-posterior for CIFAR10



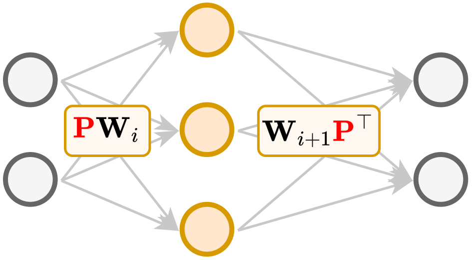
Ainsworth, Samuel, Hayase, Jonathan, and Srinivasa, Siddhartha. 2023.
Entezari, Rahim et al. 2022.

Building symmetries with weight permutations

STELLANTIS

- → Given $\boldsymbol{\theta}$ and $\boldsymbol{P}$, build $\boldsymbol{\theta}'$ as in the figure
- → Given q_1, define $P_{\#}q_1$ the push-forward distribution for $\boldsymbol{\theta}'$
- → By construction, $P_{\#}q_1$ is functionally equivalent to q_1

$$q(\boldsymbol{f}(\boldsymbol{\theta}, \cdot)) = q(\boldsymbol{f}(\boldsymbol{\theta}', \cdot)) \,. \tag{1}$$



Finding symmetries by looking at permutations

STELLANTIS

Assume two independently trained VI solutions q_0 and q_1

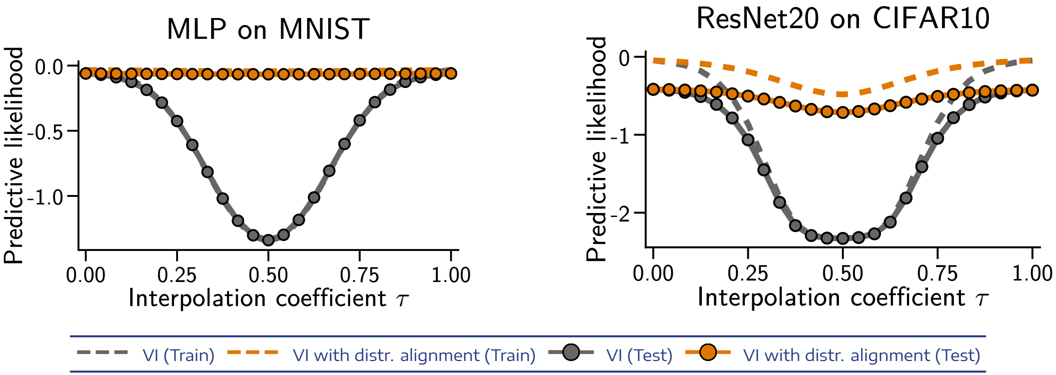
Objective

Given q_0 and q_1, find P s.t. $P_{\#}q_1$, functionally equivalent to q_1, is aligned to q_0.

$$\underset{\boldsymbol{P}\in\mathbb{S}(d)}{\arg\min}\, \mathcal{W}_2^2(P_{\#}q_1, q_0) = \underset{\{P_i\}}{\arg\min} \mathcal{W}_2^2\left(P_{1\#}q_1^{(1)}, q_0^{(1)}\right) + \mathcal{W}_2^2\left(\left(P_2 \circ P_1^\top\right)_{\#} q_1^{(2)}, q_0^{(2)}\right) + \cdots + \mathcal{W}_2^2\left(\left(P_{L-1}^\top\right)_{\#} q_1^{(L)}, q_0^{(L)}\right),$$

Solution: We approximate the optimization with a coordinate descent algorithm that converges to a local minimum of the Wasserstein distance.

Low barrier solutions



- → Loss barriers always appear between two solutions in the standard VI approach
- → With alignment we can find solutions with zero loss barrier for MLPs and nearly-zero loss barrier for ResNet20.

On permutation symmetries in Bayesian neural network posteriors: a variational perspective

Simone Rossi Stellantis, France
Ankit Singh Stellantis, India
Thomas Hannagan Stellantis, France

Abstract

The elusive nature of gradient-based optimization in neural networks is tied to their loss landscape geometry, which is poorly understood. However recent work has brought solid evidence that there is essentially no loss barrier between the local solutions of gradient descent, once accounting for weight-permutations that leave the network's computation unchanged. This raises questions for approximate inference in Bayesian neural networks (BNNs), where we are interested in marginalizing over multiple points in the loss landscape. In this work, we first extend the formalism of marginalized loss barrier and solution interpolation to BNNs, before proposing a matching algorithm to search for linearly connected solutions. This is achieved by aligning the distributions of two independent approximate Bayesian solutions with respect to permutation matrices. We build on the results of Ainsworth et al. (2023), reframing the problem as a combinatorial optimization one, using an approximation to the sum of bilinear assignment problem. We then experiment on a variety of architectures and datasets, finding nearly zero marginalized loss barriers for linearly connected solutions.

1 Introduction

Throughout the last decade, deep neural networks (DNNs) have achieved significant success in a wide range of practical applications, becoming the fundamental ingredient for e.g., computer vision [e.g., 57, 25, 41, 64], language models [e.g., 24, 82, 17] and generative models [e.g., 54, 96, 97, 98, 102, 34]. Despite recent important advancements, understanding the loss landscape of DNNs is still challenging. The characterization of its highly non-convex nature, its relation with architectural choices like depth and width and the connection with optimization and generalization are just some of the problems which have been the focus of extensive research in the last few years [e.g., 76, 26, 36, 32, 2, 30, 38, 79]. It is well known, for example, that one of the fundamental characteristics of deep neural networks is their ability to learn hierarchical features, and in this regards deeper networks seem to be exponentially more expressive than shallower models [e.g., 3, 5, 7, 8, 18, 109], leading the loss landscape to have many optima due to symmetries and over-parameterization [76, 26, 111, 94]. At the same time, the role of the depth of a model in relation with its width is far less understood [80], despite wide neural networks exhibiting important theoretical properties in their infinite limit behavior [e.g., 73, 23, 29, 53, 48, 37, 20].

Two notions that have been useful to shed light on the geometry of loss landscapes are that of *loss barriers* and *mode connectivity* [36, 26]. The mode connectivity hypothesis states that given two points in the landscape, there exists a path connecting them such that the loss is constant or near constant (or, said differently, the loss barrier is null). We refer to *linear mode connectivity* when the path connecting the two solutions is linear [32]. Recently, evidence has surfaced that stochastic gradient descent (SGD) solutions to the loss minimization problem can be linearly connected. Indeed, Entezari et al. [30] discuss the role of permutation symmetries from a loss connectivity viewpoint, conjecturing the possibility that mode connectivity is actually linear once accounting for all permutation invariances.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

On permutation symmetries in Bayesian neural network posteriors: a variational perspective

Follow the QR code for the poster schedule and location

References I

- Ainsworth, Samuel, Jonathan Hayase, and Siddhartha Srinivasa (2023). "Git Re-Basin: Merging Models modulo Permutation Symmetries". In: The Eleventh International Conference on Learning Representations.
- Entezari, Rahim et al. (2022). "The Role of Permutation Invariance in Linear Mode Connectivity of Neural Networks". In: International Conference on Learning Representations.